

# TASKS & HYPOTHESES

---

## 昨天总结

---

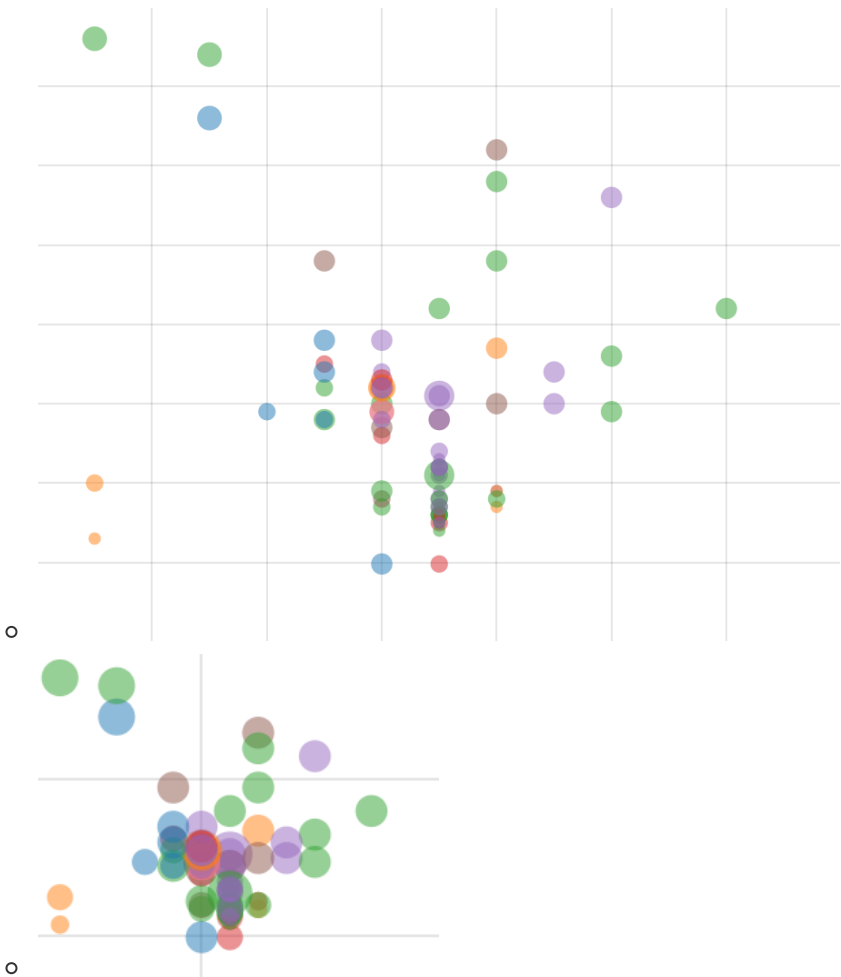
- 密度相关evaluation的整理总结
  - 怎么提现我们的不同，要多思考，论文需要claim
  - 注意笔记，思考，细化
- 基于此的一些猜想
  - 猜想，还比较混乱，逻辑不是很清晰，看起来费劲，继续细化
  - 现在你的hy，写的好技术，正经写，是很simple，常识性的表达的，注意哈。建议你早点这样写，你现在那些技术，可以放在hy的解释，或者干脆是我们内部预测的内容
  - 要多站读者，用户角度想，忌讳过多过早，细节化，技术化

## 思路回顾

---

### 场景

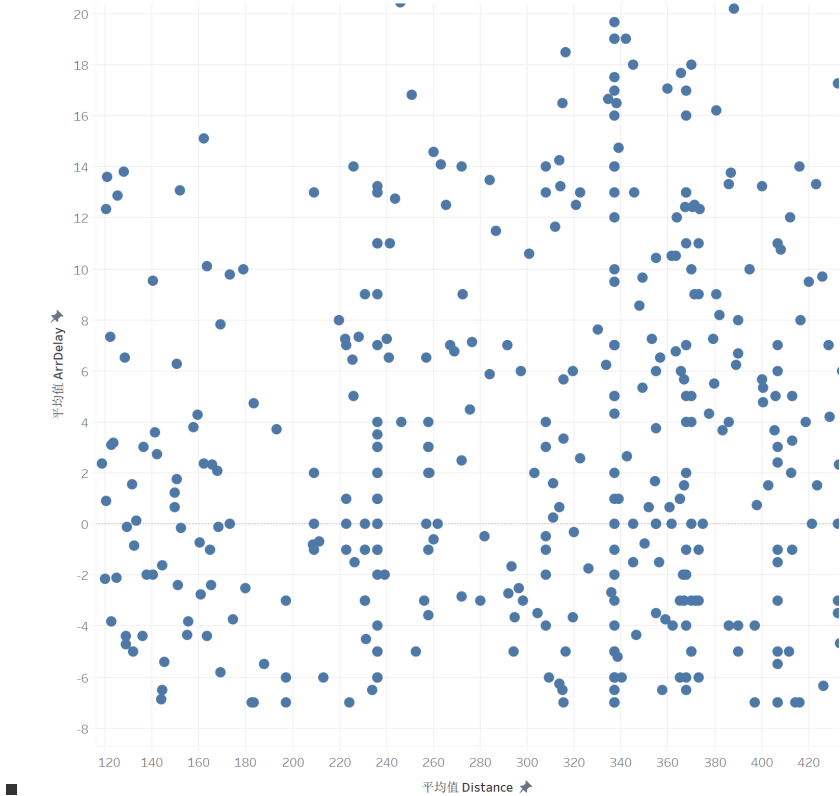
- 散点图尺寸变化
- 在不同设备之间转移
- 交互产生的变化
  - small multiple -> detailed
  - zoom in
- 坐标轴缩放样例

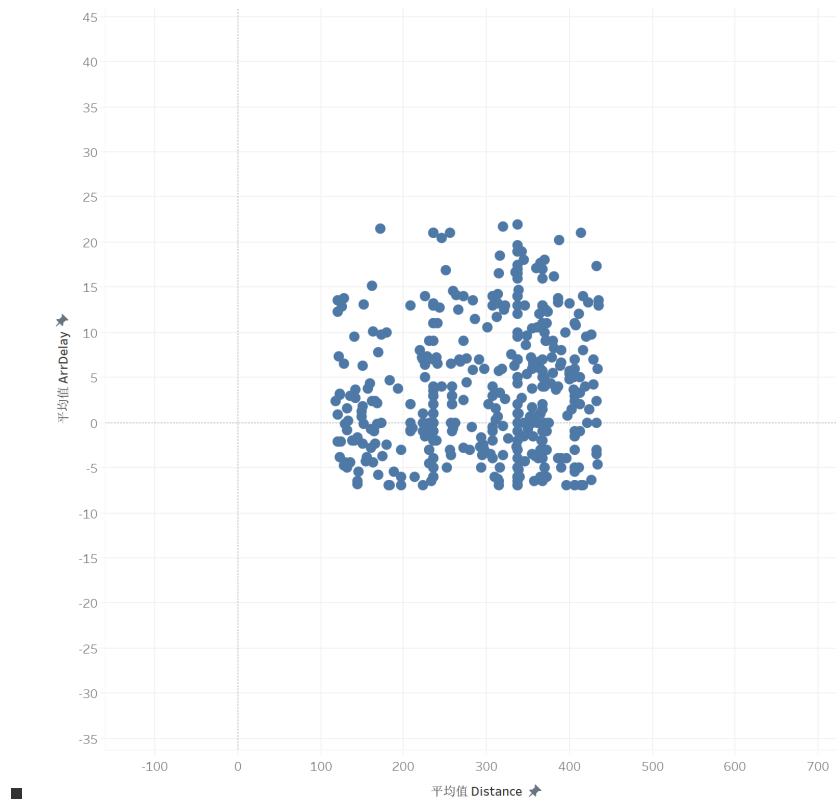


感官上还是有很明显的不同的: 纵向的直线(也就是横轴是离散的)不易被察觉

• 整个Tableau Education要耍

- tableau 没有等比缩放
- S1 缩小后更容易察觉竖直连线





## 期望得到的结果

- 之前写的
  - 等比缩放/坐标轴缩放各自会产生一定的失真(但不一样)
  - 对产生失真的原因有一定的解释
    - density和numerosity的感知模式
    - 低层->中层
  - 提出一些解决方法
    - 改变半径
- 读者希望从论文中读到些什么?
  - 这样缩放有什么问题么?
  - 那怎样的缩放是好的?
  - 等比缩放还有适用场景么?

## 论文阅读

- [1] "Visual quality metrics and human perception : an initial study on 2D projections of large multidimensional data"
  - 做cluster/class区分度的measure
  - 总结了 [2] [3] 中的一些measure
  - 一个简单的用户实验 (class区分度)

- [2] "Combining automated analysis and visualization techniques for effective exploration of high dimensional data"
  - Class Density Measure (CDM) and Histogram Density Measure (HDM)
    - identify those plots that show minimal overlap between the classes
- [3] "Selecting good views of high-dimensional data using class consistency"
  - Class Consistency Measure (CCM)
    - based on the distance of data points to their cluster centroid
  - Histogram Density Measure (HDM)
    - considers the class distribution of the points in the 2D scatter plot when they are projected on the axes
    - the measure is based on the analysis of the amount of overlap among points in the same histogram bin
- [4] "Quality metrics in high-dimensional data visualization: An overview and systematization"
  - 主体内容和perception无关
  - All the metrics that work in the image space try to simulate the human pattern recognition machinery to some extend.
  - needs a much deeper investigation: first step in this direction [1]
  - it is necessary to validate and tune the image space metrics in a way that the parameters take models of human perception into account ["Judging correlation from scatterplots and parallel coordinate plots", ...]
- [6] "A Taxonomy of Visual Cluster Separation Factors"
  - 对于cluster的评判似乎都是建立在已经有了颜色标注的情况下
  - 这些轴可以作为参考, 思考那些情况会影响计算和人类判断

- 可以考虑一下这些轴(包括scagnostics的)在缩放的情况下是否会改变

1340

Sedlmair et al. / A Taxonomy of Visual Cluster Separation Factors

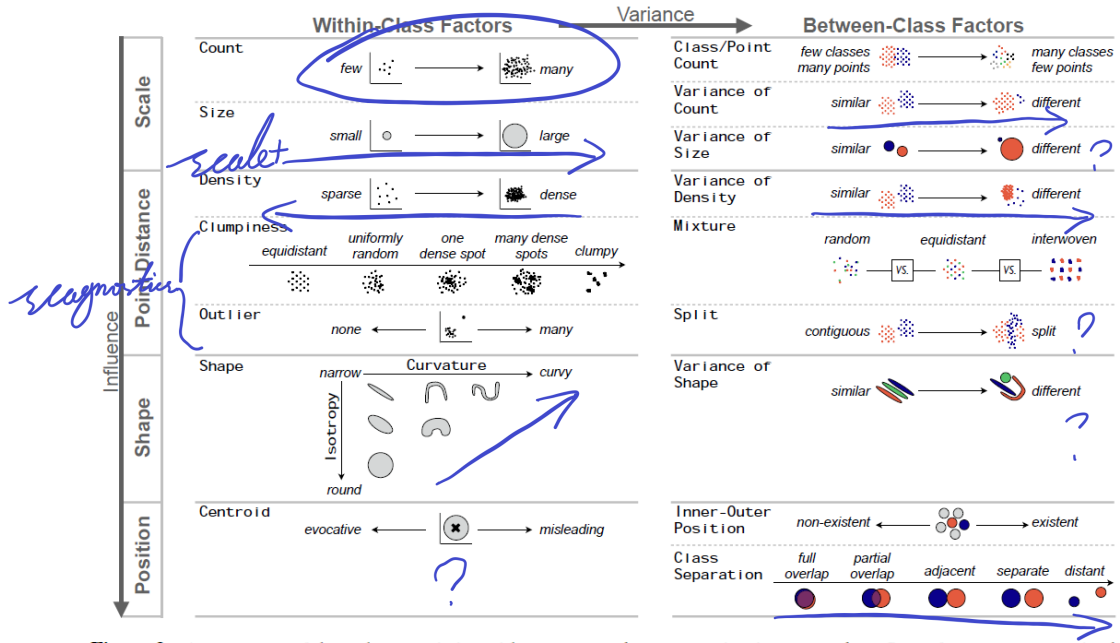
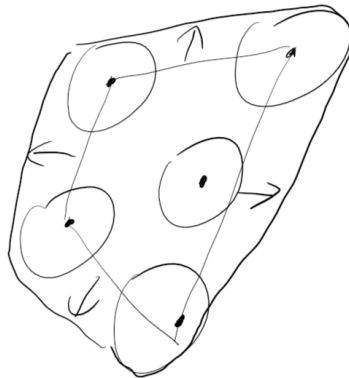


Figure 3: A taxonomy of data characteristics with respect to class separation in scatterplots. Some factors are organized as axes (arrows) while others are binned. Between-Class factors often result from the variance of Within-Class factors (horizontal dependencies), and factors at the top can strongly influence factors below them (vertical dependencies). Class Separation is therefore dependent on all other factors.

- scagnostics 全是无量纲的 与具体绘制方法无关的, 不受缩放影响
- 不过其中的一些参数可能可以有关? 例如outlier中分位点的选择, 或者convex hull可



以往外扩展一个半径?

- 数据集很多, 包括真实的和生成的, 如有需要可以参考

- outlier做的很少, perception相关全无

## 新的尝试

### idea

- 这些感知应当都与距离的判断, 或者说occupancy有关
  - density/numerosity: occupancy of points
  - cluster: 相近点互相吸附
  - outlier: cluster -> 剩余点/过小的cluster

- correlation: 这个可能不太一样
- pattern: 类似上文S1中的情况
- 基于此, scale/point r 会影响occupancy
- 缩放后, 一些measure(e.g. scagnostics)会改变么
  - 见上节 论文阅读 >> [6]
- 可能还是想复杂了, 只需要抓住几个重点就行了
  - 等比缩放/坐标轴缩放
  - 什么是失真/会发生什么样的失真
- 与之前工作的区别?
  - 更系统
    - 除了density/numerosity还包括了cluster/correlation/outlier/...
  - 更贴近特定可视化场景
    - density部分和那几篇 density/numerosity X size 的区别在哪?
    - 他们都没有等比缩放(点大小不变), 类似于坐标轴缩放

## new hypotheses

整理语言, 给读者看, 更简洁易懂, 不含具体技术细节

- 缩放比与各种属性的感知的关系, 存在一个拐点/极值点, i.e. 最适合的缩放比
- 这个缩放比对各个感知属性
  - (branch 1) 是一致的
  - (branch 2) 不一致; 对于density和numerosity不一样, 其他的介于两者之间(是两者的组合)

## 今日总结

---

### 8.3 总结

- 调研了其他散点图相关感知的文章, 包括 outlier detection 和 cluster separation
  - 总的来说, 做感知相关的比较少, 大部分都是做 quality metrics 的
  - 少数几篇做了感知研究的, 主要focus在对 quality metrics 的有效性的验证上
  - 目前没有看到与我们实验目标有重合的研究
- 总结目前的想法
  - 场景: 散点图尺寸缩放时会发生失真
    - a. 从显示器投影到大屏或者手机
    - b. 从small multiple选择小图放大仔细查看
  - 失真: 观察缩放后的视图会产生与原始视图不同的感知结果
    - 对于没有观察到原始视图的人可能会更明显 (场景1)
    - 看到原始视图可以在一定程度上修正失真 (场景2)
  - hypotheses
    - a. 等比缩放会导致对密度的感知发生偏移, 缩小时密度感知偏低, 放大时则相反

- b. 这种偏移会影响到用户对散点图的其他类型感知, 包括 outlier detection, cluster separation等
- c. correlation的感知相对独立, 不会受到这种效应影响
- d. 在缩小散点图时适当增加半径, 或者在放大散点图时适当减少半径, 可以一定程度上补偿发生的感知偏移
- e. 如果用户可以获得关于密度和尺寸变化的额外线索, 可以在一定程度上补偿尺寸变化产生的感知偏移
- f. 用户对散点图的感知是分区的, 对局部的感知可以一定程度上提供密度和尺寸变化的线索, 从而减轻感知的偏移 (H5)
- o tasks
  - a. 测试等比缩放时, 用户对 density, numerosity, outlier detection, cluster separation 和 correlation 的感知变化 (H1-3)
  - b. 测试密度变化对 outlier detection, cluster separation 和 correlation 的感知变化 (H2-3)
  - c. 测试半径以不同于视图整体的缩放比例进行变化时, 对密度感知的影响 (H5)
  - 3b. 测试半径以不同于视图整体的缩放比例进行变化时, 对其他感知的影响 (H5)
  - d. ? (H5)
  - e. 将全图划分区域进行T1/T2 (H6)

### 8.3 总结

1. 整理了一下近期工作进展和安排, 准备全体大会(然而没有开)
2. 今天主要看density/numerosity相关的文章 (预计明天会看cluster/outlier相关); 总结了一下
  - a. 近两年主流的认知是 低密度直接感知numerosity, 高密度为density和size共同作用 (numerosity)
  - b. 在低密度区间(直接感知numerosity)相比而言受到size和density影响较小, 结果更接近真实
  - c. 给出一定的 density和size 的 cue 可以显著提高对 numerosity 的感知
  - d. 没有见到工作给出类似 correlation 的 physical X subjective 的曲线, 可能是因为density
3. 此外, 玩耍了一下之前做的测试系统, 还在手机上尝试了一下
  - a. 发现一个问题是在PPI大于一定值的手机上的px数值显示为x2的, 需要注意
4. 基于此, 做了一些猜想
  - a. numerosity/density perception X scale 的函数曲线分3段, 直线(接近斜45°)-递减-直线(
    - i. 事实上scale变化应当在某个程度上和density等效
    - ii. 等比缩放相当于改变以点个数计算的density
    - iii. 坐标轴缩放相当于改变以面积缩放的density
  - b. 对散点图的感知是分区的/局部pattern会作为cue

- i. 例如正态分布，中心区域密度较大时产生较大感知偏差，但用户可能可以通过观察边缘小密度
- ii. -> Task: 分区观察/选择
- iii. 结合上一个猜想，高斯分布和均匀分布可能会得到不同的实验结果
- c. 给用户对于尺寸变化的额外提示可以增强用户的感知准确率
  - i. 这个结论在density/numerosity上被证实
  - ii. 是否在其他visual perception上也会有类似效应?
- d. 基于上一个假设，实验顺序(先看哪个/哪个作为reference)会产生一定的影响
  - i. 这个在一些correlation的测试中有所体现[The perception of correlation in scatter plots]
  - ii. 他们的测试方法考虑了这个效应，可以作为参考